Statistical Machine Learning Techniques for Predicting Lung Cancer Prevalence

Executive summary

Zhuoran Wang, Adekunle Ajiboye, Fekadu Bayisa

Data Science Bootcamp Trainees, Erdős Institute

April 20, 2025

Introduction: Lung cancer is one of the most critical and pressing public health challenges. This study analyzes and models the prevalence of lung cancer using a combination of demographic, behavioural, and socioeconomic features.

Goal: To discern demographic, behavioral, socioeconomic and environmental drivers of lung cancer incidence in Virginia counties (2014-2018) and build predictive models to inform public health policy.

Data overview: County-level aggregated counts of lung cancer cases among populations 18 and older, with data derived from the following domains: *Demographic* (% Male, % Female, % Black, % White, % Hispanic, % aged 65+); *Behavioral* (Prevalence of Smoking, Binge Drinking, and Obesity), *Socioeconomic* (% Poverty rate, Social Deprivation Index (SDI), Median Household Income), and *Environmental* (PM2.5 Air Quality ($\mu g/m^3$)). We preprocessed the data by removing missing values, performing feature engineering, and scaling the features to better capture the features that drive cancer incidence.

Modelling Approach: Poisson generalized linear model with elastic net, which combines L_1 and L_2 penalties, is used for feature selection. XGBoost with a Poisson loss function, a boosted tree model tuned for count data, is proposed to improve predictive accuracy and identify features that drive lung cancer prevalence. After splitting the processed data into training (70%) and testing (30%) sets, both models are trained on the training data with five-fold cross-validation, and their hyperparameters are tuned using mean absolute error.

Implementation: We used several Python packages to preprocess the data and train the models. For XGBoost, we incorporated a custom loss function into the existing implementation to train the proposed model tailored to our problem.

Results: The results from the **Poisson generalized linear model with elastic net regularization** suggest strong positive associations between lung cancer prevalence and features such as smoking rates, the proportion of the population aged 65 and older, and the proportions of older Black and White individuals in the population under study. In contrast, negative associations are observed with poverty rates and the proportion of Hispanic individuals. Using **XGBoost with a Poisson loss function**, we achieved improved predictive performance, with a mean absolute error (MAE) of 5.963 compared to 6.313 from the GLM, representing a 5.44% improvement. The most important features identified by the XGBoost model include median household income, PM2.5 levels, smoking rates, and obesity prevalence. These features consistently demonstrated strong predictive power, underscoring their central role in explaining variation in lung cancer prevalence.

Conclusion: Lung cancer prevalence is driven by a combination of demographic, behavioural, and environmental features. Structural inequality and public health behaviours, such as smoking, are key contributors to cancer incidence. Machine learning techniques, when combined with domain knowledge, enhance the predictive accuracy of cancer prevalence.

Recommendation: To reduce lung cancer incidence, it is essential to prioritize interventions for high-risk groups, such as older adults, smokers, and communities facing environmental and economic stress. Besides, integrating behavioural and environmental health data into cancer prevention programs is crucial.